BALTIC JOURNAL OF LAW & POLITICS
A Journal of Vytautas Magnus University
VOLUME 15, NUMBER 4 (2022)
ISSN 2029-0454

Cite: *Baltic Journal of Law & Politics* 15:4 (2022): 416-423
DOI:10.2478/bjlp-2022-004044

Detection of Malware Attacks Using Naive Bayes Algorithm Comparing Logistic Regression Algorithm to have Improved Accuracy Rate

Borra Madhan Mohan Reddy
Research Scholar, Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical And Technical Sciences, Saveetha University, Chennai, Tamil Nadu, India: 602105.

P.Sriramya
Project Guide, Corresponding Author, Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical And Technical Sciences, Saveetha University, Chennai, Tamil Nadu, India: 602105.

Received: August 8, 2022; reviews: 2; accepted: November 29, 2022.

Abstract

Aim: To enhance the accuracy in Detection of Malware in Detection of Malware attacks Using Naive Bayes Algorithm comparing Logistic Regression Algorithm to have improved accuracy rate. **Materials and Methods:** This study contains 2 groups i.e novel Naive Bayes Algorithm (NB) comparing Logistic Regression Algorithm (LR). Each group consists of a sample size of 30. Their accuracies are compared with each other using different sample sizes also. The G-Power in the test set will be at 80%. **Results:** Data is trained in the given model so that Machine learning can function effectively. The Logistic Regression Algorithm is 50% more accurate than the Naive Bayes Algorithm of 62.8% in classifying the malware Detection.The outcomes have been acquired with a stage of importance fee of p=0.053, with a pretest power value of 80% using SPSS tools. **Conclusion:** Through this, Prediction is done for The Naive Bayes model is significantly better than the Logistic Regression in identifying Novel Malware Attacks Analysis. Naive Bayes can be also considered as a better option for the classification of Novel Malware Attacks Analysis.

Keywords
Malware Detection, Machine Learning, Novel Malware Attacks Analysis, Naive Bayes Algorithm, Logistic Regression Algorithm.

INTRODUCTION

Detecting malicious detection is a difficult task. The huge, ever-growing ecosystem of malicious software and tools presents a daunting challenge for network operators and IT administrators. One of the most extensively used methods for detecting and preventing malicious and unwanted software is antivirus software (Win, Tianfield, and Mair 2015). However, the increasing sophistication of modern malicious software means that it is increasingly challenging for any single vendor to develop signatures for every new threat. Indeed, a recent Microsoft survey found more than 45,000 new variants of backdoors, Trojans, and bots during the second half of 2006 (Christodorescu et al. 2007). This document advocates a replacement model for host malware detection based on the implementation of antivirus as a network service in the cloud (Likarish, Jung, and Jo 2009).

This mannequin permits the identification of malicious and undesirable software by using more than one detection engine severally (Watson et al. 2016).

Most cited articles, The websites visited reference are IEEE and Google Scholar. IEEE has 90 citations and Google scholar has about 170 citations. “A survey of malware detection techniques” (Win, Tianfield, and Mair 2015) has been cited by 161, "Efficient Detection of Zero-day Android Malware Using Normalized Bernoulli Naive Bayes" (Sayfullina et al. 2015) was Cited by 15, "Malware: An Overview on Threats, Detection and Evasion Attacks" was Cited by 22. This paper combines detection techniques, static signature analysis and dynamic evaluation detection. Using this mechanism, we find that Novel Cloud Malware discovery affords 35% higher detection coverage against the latest threats compared in accordance with an individual antivirus machine then a 98% discovery dimension throughout the Malware analysis (Salam, Maged, and Mahmoud 2014). Malware safety of pc structures is a totally crucial assignment in Cyber-Security (McDole et al. 2020). Even one unmarried assault is enough to lose our data (Nancy et al. 2016).

Previously our team has a rich experience in working on various research projects across multiple disciplines (Venu and Appavu 2021; Gudipaneni et al. 2020; Sivasamy, Venugopal, and Espinoza-González 2020; Sathish et al. 2020; Reddy et al. 2020; Sathish and Karthick 2020; Benin et al. 2020; Nalini, Selvaraj, and Kumar 2020).This paper combines detection techniques, static signature analysis and dynamic evaluation detection. The malware detection accuracy of the unique algorithm and hybrid model is compared, and the result shows that the hybrid system is better on Novel Malware Attacks Analysis (Santoso et al. 2019). Based on the literature survey, we have a thorough understanding of state of the art in malware detection and noted that malware detection requires a highly accurate and better perfuming model for the malware to be detected with higher precision the problem to be minimized. Different sorts of malwares were taken for evaluation and additionally in comparison 2 algorithms — NB, LR

MATERIALS AND METHODS

The research work was performed in the Data analytics Lab in the Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical And Technical Sciences. The sample size taken for conducting the experiment was 10. Two groups are considered as classifiers algorithms in order to classify prediction of fare amount, machine learning classification algorithms are used. The work was carried out on 100000 records from a data-master dataset. (Setyawan, Awangga, and Efendi 2018) The accuracy in classifying the malware was performed by evaluating two groups namely Naive Bayes Algorithm and Logistic Regression. A total of 10 iterations were performed on each group to achieve better accuracy. The sample sizes of both groups are 30% and 70% total sample sizes taken are 100% of the data. The same set of sample sizes will be used for both Algorithms. Iteration-1 for the Train set and Iteration-2 for the Test set will have 80% of the G-Power. (Setyawan, Awangga, and Efendi 2018; Pranckevičius and Marcinkevičius 2017)This helps to create a more Accurate Prediction for the Novel Malware Attacks Analysis using Machine Learning models.

Naive Bayes(NB) algorithm

Naive Bayes is a probabilistic machine learning algorithm that can be utilized in a wide assortment of grouping tasks. The name naive is utilized on the grounds that it accepts the provisions that go into the model are free of one another. Numerically Given the Bayesian calculation is addressing a class variable and the arrangement of qualities are, Conditional probability of A given B can be registered as:

$$P(A \mid B) = P(A \cap B) / P(B) \qquad (1)$$

Logistic Regression

Logistic Regression is a classification algorithm for categorical variables like Yes/No, True/False, 0/1, etc,. Logistic regression transforms its product using the logistic sigmoid function to return a chance value. The definition of the logistic function is given in equation (2)

$$\sigma(t) = 1/1{+}\mathrm{et} \qquad (2)$$

Equation (3) function is used to transform the typical linear regression formula

$$f(x) = \beta 0 + \beta 1x \qquad (3)$$

The resulting equation is shown in Equation 4. In this formula, p(x) represents the probability that an input sample belongs to the target 1. That is, the probability that an application is malicious given that it is making the observed system calls.

$$p(x) = 1/1 + e{-}\beta 0 - \beta 1x^{\blacksquare\blacksquare} \qquad (4)$$

Logarithmic transformation on the outcome variable allows us to model a non-linear association in a linear way. This is the equation used in Logistic Regression. Logistic regression is a supervised learning algorithm used to predict a dependent categorical target variable. In essence, if you have a large set of data that you want to categorize, logistic regression may be able to help.

Statistical Analysis

The SPSS (Statistical Package for the Social Sciences) statistical software was used in the research for statistical analysis. Group statistics and independent sample t-tests were performed on the experimental results and the graph was built for two groups with two parameters under study. The Statistical Comparison of Novel Malware Attacks Analysis using two Sample groups was done with the SPSS Version 25. The Analysis was done using the Mean, Median, Independent T-Test, and Deviation. For each sample size of data, the Accuracy is deviating between 3% to 5 %. So we finally sent all the Test sizes and also their Accuracy into the Spss tool and found the Average Accuracy values of the Naive Bayes Algorithm Classifier and the Logistic Regression Algorithm Classifier.

RESULTS

The proposed algorithm Naive Bayes and existing algorithm Logistic Regression (LR) algorithm were run at a time in an Anaconda-Jupyter. Fig. 1 shows an Architecture diagram for malware classification. As the sample sets are executed for a number of iterations the accuracy values of Naive Bayes(NB) and Logistic Regression(LR) Algorithm classifiers vary for the classification of accuracy shown in Table 1.

The observed values for the metrics of Group Statistics, the mean accuracy, and the standard deviation for the Naive Bayes(NB) Algorithm are 62.2 and 0.37014. The Logistic Regression(LR) Algorithm's mean accuracy is 49.92 and the standard deviation is 0.66106. The Naive Bayes(NB) Algorithm also obtained a standard error mean rate of 0.16553 whereas the Logistic Regression(LR) Algorithm obtained an error mean rate of 0.27563 as shown in table 2.

Analysis of the overall classification of Detection of Malware in Cloud storage Data by Naive Bayes and Logistic Regression Algorithm models shows the classification of the detecting malware. Naive Bayes (62.7%) shows better accuracy than Logistic Regression (50%). Statistical Analysis of Mean, Standard deviation and Standard Error and Accuracy of Naive Bayes and Logistic Regression Algorithm is done.

Then an independent sample test of 5 samples was performed, Naive Bayes Algorithm obtained a mean difference of 12.01 and a standard error difference of 0.33882. When compared to other algorithm performance, the Naive Bayes Algorithm performed better than the Logistic Regression Algorithm and the significance value of 0.053 ($p<0.05$) shows that our hypothesis is valid as given in Table 3. There is a statistically significant difference in Accuracy values between the algorithms. Logistic Regression had obtained higher accuracy compared to Naive Bayes. Fig. 2 represents the bar chart of accuracies with standard deviation error is plotted for both the algorithms.

DISCUSSION

The Naive Bayes and Logistic Regression algorithm classifiers on a dataset acquired from diverse sources like Kaggle, Github, et al. are compared during this section. After completing preprocessing and extraction on the dataset, the dataset was separated into portions for training and testing. The accuracy is calculated using both Naive Bayes and Logistic Regression (LR) Algorithm. The Naive Bayes Algorithm was better than Logistic Regression in every way. The accuracy of a classifier is critical in determining the efficacy of Novel Malware Attacks Analysis in Cloud storage to reduce false detection.

In the proposed system, we have used traditional detection techniques (optimizing pattern) as per static signatures and dynamic detection Malware Detection ("Optimizing Spam Detection in Twitter by Using Naïve Bayes, Logistic Regression and Stochastic Gradient Descent with Whale Optimization Algorithm and Genetic Algorithm" 2020). Then, we have chosen safer system methods as well as speed and modernity to rival existing anti-virus (Setyawan, Awangga, and Efendi 2018). The proposal of this work is to find the best solutions to the problems of anti-viruses and improve performance and find possible alternatives for a better working environment without problems with high efficiency and flexibility (Yadav 2021). We used the optimal traditional methods and modern methods to detect viruses, for unknown and already detected viruses through the Malware Detection (Hasanli and Rustamov 2019). The shares of correct predictions divided by the whole number of guesses is known as accuracy. We evaluated the accuracy of each machine learning technique to figure out which was the foremost effective.We used sci-kits sklearn. Metrics. Accuracy score to calculate the classifier accuracy for Naive Bayes and Logistic Regression (LR) algorithm.

From the database, the algorithm will get matched Detection of Malware in Cloud storage, also as basic profile information about. These findings are being provided to an interface that will display and populate a machine learning algorithm that discovers and formalizes the principles that underlie the data it sees. Despite the actual fact that the presented methodology yielded good results, the approach's shortcoming is that it needs to be enhanced to reduce false detection of malware. This may be avoided in the future by combining Naive Bayes with other approaches.

CONCLUSION

The studies on prediction are completed using the device getting to machine learning algorithms. Naive Bayes Algorithm (NB) comparing Logistic Regression Algorithm (LR) are giving the accuracy of 62.8% and 50.0% accuracy separately. The studies can be in addition prolonged with diverse datasets and diverse attributes for the ensemble of the device getting to know algorithms.

DECLARATION

Conflicts of Interest

The author declares no conflict of Interest.

Authors Contributions

Author VST was involved in data collection, data analysis, manuscript writing. Author PSR was involved in conceptualization, data validation, and critical review of manuscript.

Acknowledgement

The authors would like to express their gratitude towards Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (Formerly known as Saveetha University) for providing the necessary infrastructure to carry out this work successfully.

Funding: We thank the following organizations for providing financial support that enabled us to complete the study.

1. Qbec Infosol Pvt. Ltd., Chennai.
2. Saveetha University
3. Saveetha Institute of Medical and Technical Sciences
4. Saveetha School of Engineering

REFERENCES

Benin, S. R., S. Kannan, Renjin J. Bright, and A. Jacob Moses. 2020. "A Review on Mechanical Characterization of Polymer Matrix Composites & Its Effects Reinforced with Various Natural Fibres." *Materials Today: Proceedings* 33 (January): 798–805.

Christodorescu, Mihai, Somesh Jha, Douglas Maughan, Dawn Song, and Cliff Wang. 2007. *Malware Detection*. Springer Science & Business Media.

Gudipaneni, Ravi Kumar, Mohammad Khursheed Alam, Santosh R. Patil, and Mohmed Isaqali Karobari. 2020. "Measurement of the Maximum Occlusal Bite Force and Its Relation to the Caries Spectrum of First Permanent Molars in Early Permanent Dentition." *The Journal of Clinical Pediatric Dentistry* 44 (6): 423–28.

Hasanli, Huseyn, and Samir Rustamov. 2019. "Sentiment Analysis of Azerbaijani Twits Using Logistic Regression, Naive Bayes and SVM." *2019 IEEE 13th International Conference on Application of Information and Communication Technologies (AICT)*. https://doi.org/10.1109/aict47866.2019.8981793.

Likarish, Peter, Eunjin Jung, and Insoon Jo. 2009. "Obfuscated Malicious Javascript Detection Using Classification Techniques." *2009 4th International Conference on Malicious and Unwanted Software (MALWARE)*. https://doi.org/10.1109/malware.2009.5403020.

McDole, Andrew, Mahmoud Abdelsalam, Maanak Gupta, and Sudip Mittal. 2020. "Analyzing CNN Based Behavioural Malware Detection Techniques on Cloud IaaS." *Lecture Notes in Computer Science*. https://doi.org/10.1007/978-3-030-59635-4_5.

Nalini, Devarajan, Jayaraman Selvaraj, and Ganesan Senthil Kumar. 2020. "Herbal Nutraceuticals: Safe and Potent Therapeutics to Battle Tumor Hypoxia." *Journal of Cancer Research and Clinical Oncology* 146 (1): 1–18.

Nancy, Nancy, Sanjay Silakari, Uday Chourasia, and Uit Rgpv. 2016. "A Survey Over the Various Malware Detection Techniques Used in Cloud Computing." *International Journal of Engineering Research and*. https://doi.org/10.17577/ijertv5is020388.

"Optimizing Spam Detection in Twitter by Using Naïve Bayes, Logistic Regression and Stochastic Gradient Descent with Whale Optimization Algorithm and Genetic Algorithm." 2020. *JOURNAL OF XI'AN UNIVERSITY OF ARCHITECTURE & TECHNOLOGY*. https://doi.org/10.37896/jxat12.03/225.

Pranckevičius, Tomas, and Virginijus Marcinkevičius. 2017. "Comparison of Naive Bayes, Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression Classifiers for Text Reviews Classification." *Baltic Journal of Modern Computing*. https://doi.org/10.22364/bjmc.2017.5.2.05.

Reddy, Poornima, Jogikalmat Krithikadatta, Valarmathi Srinivasan, Sandhya Raghu, and Natanasabapathy Velumurugan. 2020. "Dental Caries Profile and Associated Risk Factors Among Adolescent School Children in an Urban South-Indian City." *Oral Health & Preventive Dentistry* 18 (1): 379–86.

Salam, Safaa, Maged, and Mahmoud. 2014. "Malware Detection in Cloud Computing." *International Journal of Advanced Computer Science and Applications*. https://doi.org/10.14569/ijacsa.2014.050427.

Santoso, Irvan, Yaya Heryadi, Harco Leslie Hendric Warnars, Lili Ayu Wulandhari, Lukas, and Edi Abdurachman. 2019. "Malware Detection Using Hybrid Autoencoder Approach for Better Security in Educational Institutions." *2019 IEEE International Conference on Engineering, Technology and Education (TALE)*. https://doi.org/10.1109/tale48000.2019.9225899.

Sathish, T., and S. Karthick. 2020. "Gravity Die Casting Based Analysis of Aluminum Alloy with AC4B Nano-Composite." *Materials Today: Proceedings* 33 (January): 2555–58.

Sathish, T., D. Bala Subramanian, R. Saravanan, and V. Dhinakaran. 2020. "Experimental Investigation of Temperature Variation on Flat Plate Collector by Using Silicon Carbide as a Nanofluid." In *PROCEEDINGS OF INTERNATIONAL CONFERENCE ON RECENT TRENDS IN MECHANICAL AND MATERIALS ENGINEERING: ICRTMME 2019*. AIP Publishing. https://doi.org/10.1063/5.0024965.

Sayfullina, Luiza, Emil Eirola, Dmitry Komashinsky, Paolo Palumbo, Yoan Miche, Amaury Lendasse, and Juha Karhunen. 2015. "Efficient Detection of Zero-Day Android Malware Using Normalized Bernoulli Naive Bayes." *2015 IEEE Trustcom/BigDataSE/ISPA*. https://doi.org/10.1109/trustcom.2015.375.

Setyawan, Muhammad Yusril Helmi, Rolly Maulana Awangga, and Safif Rafi Efendi. 2018. "Comparison Of Multinomial Naive Bayes Algorithm And Logistic Regression For Intent Classification In Chatbot." *2018 International Conference on Applied Engineering (ICAE)*. https://doi.org/10.1109/incae.2018.8579372.

Sivasamy, Ramesh, Potu Venugopal, and Rodrigo Espinoza-González. 2020. "Structure, Electronic Structure, Optical and Magnetic Studies of Double Perovskite Gd2MnFeO6 Nanoparticles: First Principle and Experimental Studies." *Materials Today Communications* 25 (December): 101603.

Venu, Harish, and Prabhu Appavu. 2021. "Experimental Studies on the Influence of Zirconium Nanoparticle on Biodiesel–diesel Fuel Blend in CI Engine." *International Journal of Ambient Energy* 42 (14): 1588–94.

Watson, Michael R., Noor-Ul-Hassan Shirazi, Angelos K. Marnerides, Andreas Mauthe, and David Hutchison. 2016. "Malware Detection in Cloud Computing Infrastructures." *IEEE Transactions on Dependable and Secure Computing*. https://doi.org/10.1109/tdsc.2015.2457918.

Win, Thu Yein, Huaglory Tianfield, and Quentin Mair. 2015. "Detection of Malware and Kernel-Level Rootkits in Cloud Computing Environments." *2015 IEEE 2nd International Conference on Cyber Security and Cloud Computing*. https://doi.org/10.1109/cscloud.2015.54.

Yadav, Mithileshkumar. 2021. "Accuracy Enhancement of Diabetic Retinopathy Detection Using Naive Bayes Algorithm." *International Journal for Research in Applied Science and Engineering Technology*. https://doi.org/10.22214/ijraset.2021.37099.

TABLES AND FIGURES

Table 1. Comparing accuracy values with the different sample sizes. It represents Detection of Novel Malware Attacks Analysis, the accuracy of Naive Bayes (62%), and the Logistic Regression algorithm (50%).

Iteration	Naive Bayes	Logistic Regression
1	62%	50.0%
2	62.5%	49.5%
3	61.5%	50.5%
4	62.3%	49.9%
5	61.9%	50.3%

Table 2. Statistical Analysis of Mean, Standard deviation and Standard Error of and Sensitivity of Naive Bayes and Logistic Regression Algorithm. There is a statistically significant difference in Accuracy values in the algorithms. Logistic Regression had the highest Accuracy (50%) and Sensitivity (62%) compared with Naive Bayes. The Standard error is also less in Naive Bayes in comparison to Logistic Regression Algorithm.

Accuracy Group	N	Mean	Std. Deviation	Std.Error Mean
Naive Bayes	5	62.2000	.37014	.16553
LR	5	49.9200	.66106	.29563

Table 3. Comparison of the significance level for Naive Bayes and Logistic Regression algorithms with value p = 0.053. Both Naive Bayes and Logistic Regression have a significance level less than 0.05 in terms of accuracy with a 95% confidence interval.

Accuracy	Levene's Test for Equality of Variances		T-test for Equality of means						
	F	Sig.	t	df	Sig(2-tailed)	Mean Difference	Std. Error Difference	95% confidence interval of the Difference	
								Lower	Upper
Equal variances assumed	.970	.053	-35.446	8	.000	-12.0100	.33882	-11.22868	-12.79132
Equal variances not assumed			-35.446	6.953	.000	-12.0100	.338822	-11.18992	-12.83008

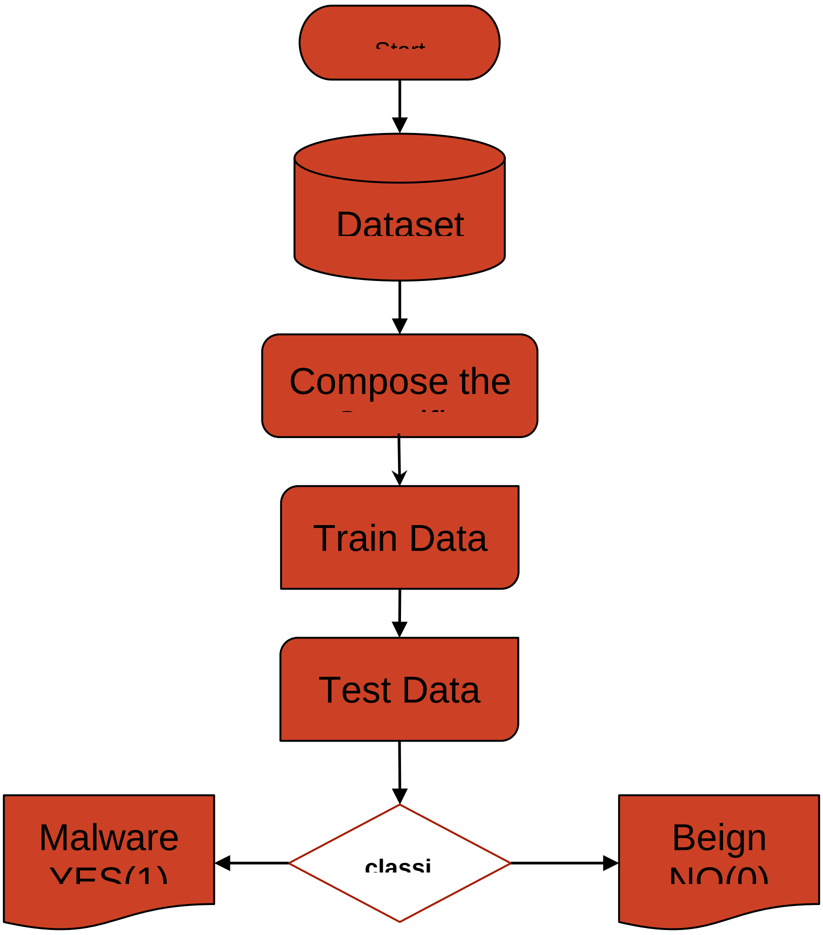

Fig. 1. Architecture Diagram for Malware Detection Analysis

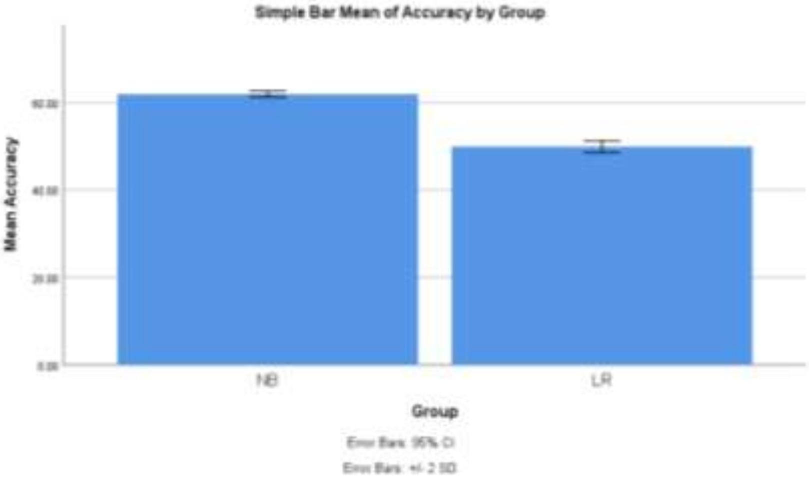

Fig 2. Bar Graph Comparison on mean accuracy of Logistic Regression with Naive Bayes Algorithm. The mean accuracy for NB is 62% and for LR is 50%. The standard errors appear to be less in Logistic Regression compared to Naive Bayes. Logistic Regression appears to produce more consistent results with higher accuracy. X-Axis: Logistic Regression vs Naive Bayes Algorithm. Y-Axis: Mean accuracy of detection +/- 2 SD, Error Bars 95% CI.